International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 2 (2018) pp. 1501-1505

Fuzzy Distance Measure Based Affinity Propagation Clustering

Omar M. al-akash[1], **Sharifah Sakinah Syed Ahmad**[2] **and Mohd Sanusi Azmi**[3]

[1,2,3] *Faculty of Information &Communication Technology, Universiti Teknikal Malaysia Melaka Hang Tuah Jaya, 71600, Durian Tunggal, Melaka, Malaysia.*
E-mail: omarakash2010@gmail.com

Abstract

Affinity Propagation (AP) is an effective algorithm that find exemplars repeatedly exchange real valued messages between pairs of data points. AP uses the similarity between data points to calculate the messages. Hence, the construction of similarity is essential in the AP algorithm. A common choice for similarity is the negative Euclidean distance. However, due to the simplicity of Euclidean distance, it cannot capture the real structure of data. Furthermore, Euclidean distance is sensitive to noise and outliers such that the performance of the AP might be degraded. Therefore, researchers have intended to utilize different similarity measures to analyse the performance of AP. nonetheless, there is still a room to enhance the performance of AP clustering. A clustering method called fuzzy based Affinity propagation (F-AP) is proposed, which is based on a fuzzy similarity measure. Experiments shows the efficiency of the proposed F-AP, experiments is performed on UCI dataset. Results shows a promising improvement on AP.

Keyword: Affinity propagation, fuzzy set, fuzzy similarity measure, clustering, unsupervised classification.

INTRODUCTION

Clustering analysis is an effective method for data mining. It aims to divide data into cluster such that members of the same cluster share higher similarities than those from different clusters [1]. Affinity Propagation (AP) is an unsupervised clustering algorithm, developed by Frey and Dueck [2] and successfully applied to broad areas of research fields. AP have been used widely in many application, [3, 4] used AP to detect activated brain regions in fMRI. choose a small number of subjects to represent the entire training data[5]. used to cluster data point in order to find optimal thresholding Levels [6]. Literature shows that AP is efficient in finding data clusters. One main advantage of AP is that it does not requires pre-specifying the number of exemplars, which is very useful in real applications.

AP identifies the exemplars among data points by simultaneously considering all data points as a potential exemplar and exchanging messages between pairs of data point until a set of exemplars found. AP uses the negative Euclidean distance as similarity metric forming a similarities matrix between data points. Moreover, AP assigns a preference value that shows how likely a data points to serve as an exemplar. AP exchanged two types of messages called "responsibility" and "availability". The method iteratively maximize the responsibility and minimize availability for each data point, until AP finds a collection of exemplars and groups the data points around these exemplars. Generally, the common choice of AP similarity is the negative Euclidean distance. Therefore, the performance of finding exemplar by AP affected by the usage of this similarity metric. Regarding to the problem that some data doesn't lie into the Euclidean space which make it hard for the Euclidean distance to capture the real relationship among data points [7]. Furthermore, the Euclidean distance is sensitive to noise in data such that AP clustering results might be degraded when data contain noise or outliers [8]. However, unless a meaningful measure of similarity between pairs of points defined, no meaningful cluster analysis is possible. Aiming to solve this knotty problem, in this paper, an improved similarity measure integrating fuzzy sets with AP proposed, which called fuzzy based AP (F-AP). Recently, many clustering methods based on the concept of fuzzy approaches have been introduces and developed to overcome the uncertainty effects based on the concept of Fuzzy set. Fuzzy set used to calculate the degree of membership of a pattern to a class according to the use of a membership function. The proposed F-AP method considers that all data points can be equally suitable as initial exemplars. A fuzzy similarity measure is utilized to calculate the distance between data points, and then, construct a similarity matrix; after that, performs the AP algorithm over the matrix to obtain exemplars.

RELATED WORK

In recent years, AP clustering proposed by Frey and Dueck [2]. Since then, numerous improvements been proposed by different researches to enhance the AP clustering efficiency. Some researchers propose a modification on the AP procedure [9-13], and some have utilizes different similarity measure and preference value calculation technique to improve its efficacy.

AP uses the negative Euclidean Distance regarding to its simplicity. However, this distance suffers from a high sensitivity even to small deformation [8]. Several similarity measure were proposed to improve the AP in the literature, the improvement aims to replace the standard Euclidean distance as it not very suitable for some kind of data like in [14] a fuzzy statistical similarity measure (FSS) is developed in evaluating the similarity between two pixel vectors. While [15] extend the generic definition of the similarity measure based on the cosine coefficient by introducing an asymmetric similarity measurement. In [16] a kullback-leibeler divergence is used for construction of the similarity matrix. In [17] The similarity between two consensus matrices is defined as the Manhattan distance. In [18] a similarity is set to the negative generalized likelihood ratio (GLR). In [19] Jaccard similarity have been used to construct the similarity matrix. [20] propose a novel

similarity measure for mixed type datasets and an adaptive AP clustering algorithm proposed to cluster the mixed datasets. The similarities are set to a negative squared error to coordinate the input of AP algorithm. [21] Proposed algorithm called Phrase affinity clustering (PAC). PAC first finds the phrase by ukkonen suffix tree construction algorithm, second finds the vector space model VSD using tf-idf weighting scheme of phrase. Third calculate the similarity matrix form VSD using cosine similarity, and then AP algorithm generate the clusters. [12] propose Multi-Exemplar Affinity Propagation MEAP that take advantage of the sparsity in data. Similarity measure defined by the number of significant feature matches normalized by subtracting means across both dimensions. [22] It finds a transformation matrix of the feature space using equivalence constraints. This letter explores this idea for constructing a feature metric (FM) and develops a novel semi-supervised feature-selection technique for hyperspectral image classification. Two feature measures referred to as band correlation metric (BCM) and band separability metric (BSM) are derived for the FM. The BCM can measure the spectral correlation among the bands, while the BSM can assess the class discrimination capability of a single band. The proposed feature-metric-based AP (FM-AP) technique utilizes AP, to group bands from original spectral channels with the FM. [23] a novel semi supervised clustering technique incremental and decremental affinity propagation (ID-AP) that incorporates labeled exemplars into the AP algorithm. Unlike standard semi supervised clustering methods, the proposed technique improves the performance by using both the labeled samples to adjust the similarity matrix and an ID-learning principle for unlabeled data selection and useless labeled samples rejection. [24] Propose a method to select an optimal thresholding value by utilizing a novel similarity metric between the data points along the gray-level histogram of the image then using Affinity Propagation (AP) to cluster the intensities based on the geodesic distance metric. It is obvious that if the similarity matrix can accurately capture relationship among data, the affinity propagation can achieve excellent clustering.

METHODOLOGY

Affinity propagation has several advantages: speed, general applicability, and suitable for large number of clusters. AP takes as input similarity measures between pairs of data points [25]. AP operates by simultaneously considering all data points as potential exemplars and iteratively exchanging messages between data points until a good set of exemplars and cluster emerges. However, AP has two limitations: the performance of AP is based on the similarity measures, and it is hard to know what value of parameter 'preference' can yield optimal clustering solution.

In the crisp method, every given object is classified into a specific cluster, where each data point attached to only one cluster. Contrary to this, the features of the objects are vague and have some tendency to be part of other clusters. The fuzzy set theory provides a powerful tool for soft partitioning of data sets. Thus, clustering by using fuzzy concepts, called fuzzy clustering. Fuzzy clustering can more effectively reflects the real world as it obtains the degree of membership of samples belonging to each class and expresses the intermediate property of their membership. Fuzzy clustering techniques use the concept of memberships to describe the degree by which a vector belongs to a cluster. The use of memberships provides fuzzy methods with more realistic clustering than hard or crisp techniques.

A. Affinity Propagation algorithm

In this section, we briefly review the AP algorithm. AP clusters data by a collection of real-valued similarities between pairs of data points to produce a high-quality set of centers and corresponding clusters. The centers are the exemplars as described by [2]. The similarity $s(i,k)$ in AP indicates how well the data point with index k is suitable to be the exemplar for data point i. For example, in Euclidean space when the goal is to minimize squared error, each similarity is set to a negative squared error, for data point x_i and x_j should be equal with the similarity of data point x_k and x_i. When appropriate, similarities may be set by hand. If x_k is likely to be chosen as an exemplar, we can set $s(k,k)$ a larger value. There are two kinds of message exchanged between data points, the responsibility $r(i,k)$ and the availability $a(i,k)$, the responsibility sent from data point i to candidate exemplar point k, reflects the accumulated evidence for how well-suited point k is to serve as the exemplar for point i, taking into account other potential exemplars for point data i While the availability $a(i,k)$ sent from candidate exemplar point k to point i, reflects the accumulated evidence for how appropriate it would be for point i to choose point k as its exemplar, taking into account the support from other points that point should be an exemplar. The responsibilities $r(i,k)$ and availability $a(i,k)$ are computed as:

Initialization:

$$r(i,k)=s(i,k)-\max\{s(i,k')\},a(i,k)=0 \quad k'\neq k \tag{1}$$

Responsibility updates:

$$r(i,k) = s(i,k) - \max\{a(i,k')+s(i,k')\} \tag{2}$$

$$r(i,k) = s(k,k) - \max\{a(k,k')+s(k,k')\} \tag{3}$$

Availability updates:

$$a(i,k) = \min\left(0, r(k,k) + \sum_{i's.t.i'\notin\{i,k\}} \max\{0, r(i',k)\}\right) \tag{4}$$

Here $r(k,k)$ is the self-responsibility, which is set to the input preference that the point k be chosen as an exemplars. If $r(k,k)$ is negative, it means that point k is currently better suited as belonging to another exemplar rather than being an exemplar itself. While the "self-availability" $a(k,k)$ reflects accumulated evidence that point k is an exemplar, based on the positive responsibilities sent to candidate exemplar k from other points. $r(k,k)$ And $a(k,k)$ is updated as follow:

$$a(k,k) = \sum_{i's.t.i' \notin \{i,k\}} \max\{0, r(i',k)\} \quad (5)$$

Making assignments:

$$c_i^* \leftarrow \arg \max_{1 \le k \le n} \; r(i,k) + a(i,k)$$

B. Fuzzy based Affinity propagation

Clustering implies a grouping of data points in a multidimensional space. Data points belongs to a specific cluster are similar. In order to find this similarity, it is necessary to define a similarity measure. In AP, the most important factor that affect the performance of AP is the similarity measures. In other words, the similarity matrix can influence the AP to achieve excellent clustering [26]. Therefore, it is possible to use advanced notions of similarity to replace the Euclidean distance. A negative fuzzy normalized Euclidean distance measures proposed as replacement for the Negative Euclidean distance.

Fuzzy set theory is the basis in studying membership relationships from fuzziness. Fuzzy set has a membership function that assigns to each element of the universe of discourse, a number from the unit interval [0, 1] to indicate the degree of belongingness to the set under consideration. In fuzzy set theory, the membership of an element to a fuzzy set is a single value between zero and one.

In this section the negative normalized Euclidean distance [27] is proposed to replace the negative Euclidean distance used in the AP clustering method. The negative normalized Euclidean distance defined as

$$d_{ne}(A,B) = \frac{1}{n}\left(\sum_{i=1}^{n} |\mu_A(x_i) - \mu_B(x_i)|^2 \right)^{1/2}$$

Where $\mu_A(x_i)$ and $\mu_B(x_i)$ are the membership functions of A and B, respectively with the condition the

$0 \le \mu_A(x_i), \mu_B(x_i) \le 1$, for $x_i \in X, i = 1,2,\ldots,n$.

F-AP utilize the negative normalized Euclidean distance in its procedure to construct the similarity matrix. In this case the value of $s(i,k)$ is obtained by

$$s(i,k) = -\frac{1}{n}\left(\sum_{i=1}^{n} |\mu(x_i) - \mu(x_k)|^2 \right)^{1/2} \quad (6)$$

And the preference value is calculated using based on the following variation:

$$s(i,i) = pr.*\min(s(i,k)) \quad (7)$$

Where pr is a value between [2, 6], min are the minimum value of all $s(i,k),(i \neq k)$ .also based on the standard AP where the preference have been chosen to be the median of all similarities as in E.q 8:

$$s(i,i) = median(s(i,k)) \quad (8)$$

C. F-AP approach

The algorithm proposed in the following is based on fuzzy similarity measure and AP and is called F-AP. Compared with the conventional AP clustering method. It simultaneously considers all data points in the feature space to be initial clustering exemplars. The procedure associated with F-AP is as follows:

Step 1: *calculate each data point membership value*

Step 2: *construct the similarity matrix based on the E.q. (6)*

Step3: *use one of the preference value E.q (7, 8) as input of the preference value*

Step 4: *initialize the responsibility and availability.*

Step 5: *update the responsibility and availability.*

Step 6: *Identify exemplars based on the summation of responsibility and availability.*

Step 7: *Assign non-exemplar sample i to the closest cluster.*

RESULTS AND DISCUSSION

A. Data Sets Used in the Experiments

The proposed improvement has been evaluated on the IRIS dataset, Wine dataset and the Teaching Assistant Evaluation (TAE) dataset obtained from UCI the Machine Learning Repository [28] to verify the improved algorithm. Where the Wine dataset is a 3-dimensional dataset, containing 178 instances with 13 features and divided into 3 clusters. The TAE dataset contains 151 instances and every data has 5 features, the TAE dataset has 3 clusters. While the Iris dataset is consist of 150 instances and 4 features Table 1 shows the size, dimension, including the number of clusters of the use datasets.

Table 1: Details of the used datasets

Dataset	Cluster	Size	Dimension
Iris	3	150	4
Wine	3	178	13
TAE	3	151	5

B. Experiments

This section compares the clustering performance between Fuzzy AP method (F-AP) and AP algorithm (AP). The F-AP and AP use same initial lam=0.5, and AP uses a variation of preference values and maxits =500. F-AP and AP implemented respectively for comparison. Our experiments is done on UCI datasets, which are listed in table 1.

We evaluate our experiments with Fowlkes-Mallows index. Fowlkes-Mallows (FM) index is an external criterion, which is used to quantity the ability of a given method to recover the true clustering structure in a data set. FM index near 1 means that clusters are good estimates of the groups.

$$FM = \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}} \qquad (9)$$

C. Experimental results

Table 2 ~ 4 list of clustering results. In table 2, NC is the resulting number of clusters in different methods. As we know, among all the parameters of AP, the most important is the preference *P*, which directly influences the number of clusters. Note that *P* is negative. In our experiments, we select different values of *P* to deal with the datasets, including the minimum and median of similarity matrix S as in E.q.(7,8).

Table 2: Clustering results of IRIS dataset

Data Set	Preference *P*	F-AP NC	F-AP FM	AP NC	AP FM
Iris	2*Min(sim)	3	0.83	3	0.84
	3*Min(sim)	4	0.74	3	0.82
	4*Min(sim)	3	0.84	3	0.82
	5*Min(sim)	3	0.83	3	0.84
	6*Min(sim)	3	0.75	46	0.53
	Median(sim)	11	0.51	12	0.47

Table 3: Clustering results of WINE dataset

Data Set	Preference *P*	F-AP NC	F-AP FM	AP NC	AP FM
Wine	2*Min(sim)	5	0.71	25	0.50
	3*Min(sim)	3	0.86	2	0.66
	4*Min(sim)	3	0.73	22	0.46
	5*Min(sim)	3	0.80	130	0.30
	6*Min(sim)	3	0.75	1	0.58
	Median(sim)	17	0.42	12	0.31

Table 4: Clustering results of TAE dataset

Data Set	Preference *P*	F-AP NC	F-AP FM	AP NC	AP FM
TAE	2*Min(sim)	7	0.29	6	0.24
	3*Min(sim)	6	0.31	4	0.30
	4*Min(sim)	3	0.5	2	0.41
	5*Min(sim)	3	0.45	2	0.41
	6*Min(sim)	3	0.45	2	0.44
	Median(sim)	18	0.18	20	0.16

DISCUSSION

Tables 2~4 shows the results F-AP and AP clustering using negative normalized Euclidean distance and the negative Euclidean distance. Based on the experiment results presented, we conclude that the fuzzy similarity measure produces an average better clustering accuracy than the Euclidean distance in both Wine and Tae datasets, while the results for Iris dataset were almost the same as the negative Euclidean distance. The negative normalized fuzzy Euclidean distance measure achieved a higher accuracy in Wine and Tae, but in Iris is almost the same as Euclidean. However, even the fuzzy measure get the accurate number of clusters in different values of preferences but still the TAE dataset results is not high enough to be consider good. Based on that a further testing on the proposed similarity measure should take place to find the reason why the method did not score a higher accuracy.

It can be observed that the resulting numbers of clusters through AP procedure for most datasets are different from the actual numbers of datasets F-AP method can find a true number of clusters based on different preference values, however, the preference value need to be optimized to get the accurate number of clusters and accuracy. Furthermore, it can be observed that using the median of similarity value to obtain an accurate number of cluster is not necessarily true for all data sets. In general, due to fact that the data points in these datasets are overlapped and loose. Error in clustering may occur which will degrade the clustering results.

CONCLUSION

An improved AP based on Fuzzy similarity distance proposed to cluster UCI datasets. Results compared with the standard Euclidean distance AP clustering. This work illustrate and concluded that improvement of AP using the similarity matrix lead to an efficient and effective result. In addition, the preference value could control the number of clusters produced

ACKNOWLEDGMENT

The authors would like to thank the Zamalah Scheme Scholarship, Universiti Teknikal Malaysia Melaka (UTeM) for providing the scholarship for this research. Besides, thanks to Faculty of Information Technology and Communication for providing research faculties and facilities.

REFERENCES

[1] Y. Zhou and Y. Xing, "Summary of affinity propagation," in *Advanced Materials Research*, 2011, pp. 811-816.

[2] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, pp. 972-976, 2007.

[3] D. Liu, W. Lu, and N. Zhong, "Clustering of fMRI data using affinity propagation," *Brain Informatics*, pp. 399-406, 2010.

[4] J. Zhang and H. Chen, "Analysis of activity in fMRI data for multitask experimental paradigm using affinity propagation clustering," in *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, 2010, pp. 553-555.

[5] F. Shi, L. Wang, Y. Dai, J. H. Gilmore, W. Lin, and D. Shen, "Pediatric Brain Extraction Using Learning-based Meta-algorithm," *Neuroimage*, 2012.

[6] B. Foster, U. Bagci, Z. Xu, B. Dey, B. Luna, W. Bishai, S. Jain, and D. Mollura, "Affinity propagation clustering determines distributed uptake regions in PET images: A computer-aided approach for quantification of pulmonary infections in small animals," in *Society of Nuclear Medicine Annual Meeting Abstracts*, 2013, p. 313.

[7] X. Zhang and J. C. Lv, "Sparse affinity propagation for image analysis," *Journal of Software*, vol. 9, 2014.

[8] L. Wang, Y. Zhang, and J. Feng, "On the Euclidean distance of images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, pp. 1334-1339, 2005.

[9] S. K. Shrivastava, J. Rana, and R. Jain, "Fast affinity propagation clustering based on machine learning," *International Journal of Computer Science Issues*, vol. 10, pp. 302-309, 2013.

[10] B. Jia, B. Yu, Q. Wu, C. Wei, and R. Law, "Adaptive affinity propagation method based on improved cuckoo search," *Knowledge-Based Systems*, vol. 111, pp. 27-35, 2016.

[11] K. Wang, J. Zhang, D. Li, X. Zhang, and T. Guo, "Adaptive affinity propagation clustering," *arXiv preprint arXiv:0805.1096*, 2008.

[12] C.-D. Wang, J.-H. Lai, C. Y. Suen, and J.-Y. Zhu, "Multi-exemplar affinity propagation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, pp. 2223-2237, 2013.

[13] I. Givoni, C. Chung, and B. J. Frey, "Hierarchical affinity propagation," *arXiv preprint arXiv:1202.3722*, 2012.

[14] C. Yang, L. Bruzzone, F. Sun, L. Lu, R. Guan, and Y. Liang, "A fuzzy-statistics-based affinity propagation technique for clustering in multispectral images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, pp. 2647-2659, 2010.

[15] A. Akl and S. Valaee, "Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, & compressive sensing," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 2270-2273.

[16] Y. Qian, F. Yao, and S. Jia, "Band selection for hyperspectral imagery using affinity propagation," *IET Computer Vision*, vol. 3, pp. 213-222, 2009.

[17] Z. Yu, L. Li, J. Liu, J. Zhang, and G. Han, "Adaptive noise immune cluster ensemble using affinity propagation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, pp. 3176-3189, 2015.

[18] Z. Xiang, G. Jie, L. Ping, and Y. Yonghong, "A novel speaker clustering algorithm via supervised affinity propagation," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4369-4372.

[19] Y. Wang and L. Gao, "Detecting protein complexes by an improved affinity propagation algorithm in protein-protein interaction networks," *Journal of Computers*, vol. 7, pp. 1761-1768, 2012.

[20] K. Zhang and X. Gu, "An affinity propagation clustering algorithm for mixed numeric and categorical datasets," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[21] S. K. Shrivastava, J. Rana, and R. Jain, "Text document clustering based on phrase similarity using affinity propagation," *International Journal of Computer Applications*, vol. 61, 2013.

[22] C. Yang, S. Liu, L. Bruzzone, R. Guan, and P. Du, "A feature-metric-based affinity propagation technique for feature selection in hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, pp. 1152-1156, 2013.

[23] C. Yang, L. Bruzzone, R. Guan, L. Lu, and Y. Liang, "Incremental and decremental affinity propagation for semisupervised clustering in multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, pp. 1666-1679, 2013.

[24] B. Foster, U. Bagci, B. Luna, B. Dey, W. Bishai, S. Jain, Z. Xu, and D. J. Mollura, "Robust segmentation and accurate target definition for positron emission tomography images using affinity propagation," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, 2013, pp. 1461-1464.

[25] X. Liu, M. Yin, J. Luo, and W. Chen, "An improved affinity propagation clustering algorithm for large-scale data sets," in *Natural Computation (ICNC), 2013 Ninth International Conference on*, 2013, pp. 894-899.

[26] X. Hu, J. Shang, F. Gu, and Q. Han, "Improving Wi-Fi Indoor Positioning via AP Sets Similarity and Semi-Supervised Affinity Propagation Clustering," *International Journal of Distributed Sensor Networks*, vol. 2015, 2015.

[27] Z. Xu and M. Xia, "Distance and similarity measures for hesitant fuzzy sets," *Information Sciences*, vol. 181, pp. 2128-2138, 2011.

[28] M. Lichman. (2013). *{UCI} Machine Learning Repository*. Available: http://archive.ics.uci.edu/ml

[29] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53-65, 1987.